DOCUMENT RESUME

ED 126 639 EC 090 527

AUTHOR Schmelkin, Liora
TITLE Statistical Power Analysis of Research in "Exceptional Children."
PUB DATE Apr 76
NOTE 12p.; Paper presented at the Annual International Convention, The Council for Exceptional Children (54th, Chicago, Illinois, April 4-9, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Educational Research; Evaluation Methods; Exceptional Child Research; *Handicapped; *Research Methodology; *Statistical Analysis
IDENTIFIERS *Power Analysis

ABSTRACT

A statistical power analysis of research studies in special education reported in volumes 39 and 40 of Exceptional Children (1972-73 and 1973-74) was conducted. The basic concepts of a power analysis (Type I and Type II errors, and conventional effect sizes) were reviewed, and the studies evaluated for statistical power. Using the .05 level of significance, the average power to detect small effects was .11 (5% having a better than 50-50 chance to declare significant findings), for medium effects the average power was .49 (43% having a better than 50-50 chance), and for large effects the average power was .82 (76% having a better than 50-50 chance). There are several approaches to increasing power, including increasing the number of Ss, increasing the level of alpha, altering the research design, using directional tests, and using highly reliable measures. Statistical power analysis should become a part of research training and one of the criteria for the evaluation of research reports. (Author/IM)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED126639

Statistical Power Analysis of Research in Exceptional Children

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Liora Schmelkin
Department of Educational Psychology
New York University
933 Shimkin Hall
New York, N.Y. 10003

Position: Doctoral Candidate

EC090527

Abstract

In recent years some behavioral scientists have attempted to alert and sensitize researchers to the important distinction between statistical significance and meaningfulness of findings in behavioral research. Moreover, they have attempted to impress researchers with the need to consider magnitudes of effects and statistical power in the design of their studies. Despite these attempts, it appears that the vast majority of studies in the social sciences are planned, executed, and reported without any concern with issues of substantive meaningfulness and the statistical power of the tests being used.

The present paper was devoted to a statistical power analysis of research studies in special education reported in Volumes 39 and 40 of Exceptional Children (1972-73, 1973-74). After reviewing the basic concepts involved in such an analysis, namely Type I and Type II errors, and conventional effect sizes (i.e., small, medium, and large), the published research was scrutinized for statistical power. It was found that the average power to detect small effects was .11, with only 5% of the tests having a better than 50-50 chance to declare the findings as being significant at the .05 level of significance. For medium effects, the average power was .49 with 43% of the tests having a better than 50-50 chance to declare the findings as being significant. For large effects, the average power was .82 with 78% having a better than 50-50 chance.

The paper concludes with a summary and recommendations for making power analysis an integral part of the research endeavor.

Statistical Power Analysis of

Research in Exceptional Children

Despite the controversy surrounding tests of significance and despite attempts to alert researchers to the need to interpret results substantively, findings are still reported almost exclusively in terms of significance. As is well known, given a large enough sample, any finding can be declared statistically significant. Consequently, it is important to distinguish between results that are statistically significant but not substantively important and results that are important but that are declared not significant because of low power in the statistical test used.

Judging from published research findings, most behavioral researchers seem to be either unaware or not concerned with the important role of statistical power analysis in the design of research (Brewer, 1972; Cohen, 1962; Hopkins, 1973). This frequently leads to situations in which results that are substantively not meaningful are declared to be statistically significant, or ones in which meaningful results are declared to be statistically nonsignificant. Either state of affairs is an unhappy one; the first being trivial, while the second is fraught with ambiguities or perplexities. "It is unfortunate that failure to confirm hypotheses has become equated with experimental failure" (Miller & Knapp, 1971, p. 7).

Of the four factors that affect the power of statistical tests (i.e., effect size, Type I error, Type II error, number of subjects), the most important, but the one most often overlooked, is the anticipated effect size (ES). The ES is "the degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false" (Cohen, 1969, pp. 9-10). The point of departure in designing research should be the determination of

the magnitude of the expected effect. Since it is rarely the case that there are no differences between groups on any variable and since any difference, no matter how small, can be found to be significant if a large enough number of subjects (N) is used, it is incumbent on the researcher to specify what difference he will consider meaningful and to design his research so that when in fact a meaningful finding is obtained the probability of declaring it to be significant is high. The determination of the minimum ES to be considered meaningful is based, among other things, on the nature of the research in the area under investigation, the investment of effort and money, and the consequences of rejecting the null hypothesis. For example, in research on the differential effects of different remediation programs, the magnitude of the ES considered meaningful depends upon the relative efforts and costs involved in implementing each of the programs, the possible impact each program may have in the area under study, as well as on related areas. While the decision about the desired ES in a given study is best made on the basis of theoretical and practical considerations, it frequently happens that the researcher does not have the information necessary for a meaningful decision. Under such circumstances, one may resort to conventional criteria for ES. Cohen (1962, 1969), for example, proposes that for an analysis of differences between groups, mean differences of one-quarter, one-half and one standard deviation be considered small, medium and large effects respectively.

As is known, the four factors affecting statistical power are interrelated, and the selection of any three of them determines the fourth. Most researchers seem to adopt the conventional levels of significance (e.g., .05 or .01) and conduct their study with whatever number of subjects is available to them. What is called for, instead, is to specify, in addition to α, the ES and the

desired power of the test (i.e., 1-β). "The seriousness of potential error determines how much power is necessary" (Miller & Knapp, 1971, p. 8). In the event one is unable to make such a determination, it has been suggested that β be set equal to .20, so that the probability of rejecting a false null hypothesis for the ES selected is 80% (Cohen, 1969). Having selected the three factors mentioned above it is possible to calculate the sample size necessary.

In an attempt to alert researchers to these problems, several surveys have been conducted that have tried to ascertain the power of representative studies in the areas of psychology and education (Brewer, 1972; Cohen, 1962). The present investigation was designed to study the statistical power of studies in the area of special education. Specifically, Volumes 39 and 40 of Exceptional Children were reviewed and the power each had of detecting small, medium, and large ES was calculated. The findings were then compared to those reported by Cohen (1962) and Brewer (1972).

Procedures

For purposes of comparison, it was necessary to impose standard conditions on the assessment. Since there was no evidence to indicate that a level of significance was set prior to the data collection in any of the articles, the .05 level of significance was used uniformly as has been done in the surveys mentioned above. In addition, nondirectionality of hypotheses was assumed throughout. In cases where total N's were reported without a breakdown, it was assumed that there were equal n's in each group. This procedure enables one to detect the maximum power available under optimal conditions. Since no mention of ES was made in any of the articles reviewed, operational definitions of small, medium, and large effect sizes for each test were

selected following Cohen (1969). No attempt was made to consider other problems in research design.

The three statistical tests most prevalent in the two volumes under review are *t* tests, analyses of variance (*F* tests), and correlational analyses (*r*).[1] Since other tests (e.g., sign tests) occurred only once or twice, this survey is limited to the above three. Of all the articles that used statistical tests, 35 contained enough information to calculate the power of the tests for the three ES. The remaining articles either used tests other than the three under study or did not provide sufficient information. This was primarily due to two ommissions: (a) no specification as to the exact nature of the test used, and (b) insufficient information as to the number of subjects. It should be noted that this state of affairs not only prohibits post hoc power analysis, it does not permit one to adequately interpret the results of the studies in question.

Of the 35 articles, 5 employed 2 types of tests. When more than one type of test was used, separate power analysis was conducted for each. Median power for each of the distinct statistical tests in a given article was used in the tabulation of the results. While other surveys report means instead of medians, it is felt that the latter is preferable since it is not affected by extreme cases.

Results and Discussion

The power distributions for the *t*, *F*, and *r* analyses are presented in Table 1. The results can be summarized as follows: The average power for small effects across the three types of tests used was .11 with only 5% of the tests having a better than 50-50 chance of detecting such effects. For medium effects the average power was .49 with 43% having a better than 50-50 chance. For large effects, the average power was .82 with 76% having a better

than 50-50 chance of detecting effects of this magnitude.

When comparing the present survey with the surveys by Cohen in the Journal of Abnormal and Social Psychology (1962) and by Brewer in the American Educational Research Journal (1972), it should be kept in mind that in each of the surveys the problem was approached in slightly different ways. Cohen's analysis combined the different tests into one grouping which yielded a power index for each article, regardless of the analyses involved. The median power findings of his survey were .17, .46, and .89 for small, medium, and large effects respectively. Brewer presented the data classified into the various statistical tests he examined (*F*, *t*, and *r*). In contrast with the present survey, he analyzed the data without an article breakdown. Consequently, there is no way of knowing the contribution of each article to the overall index reported. The combined average power of Brewer's survey was .14, .58, and .78 for small, medium, and large ES respectively. Although direct comparison cannot be made because of the somewhat different procedures, it is interesting to note that while the three surveys deal with different content areas, their findings are generally very similar. The findings of these surveys do not portray an encouraging situation. On the average, only when large effects were being studied, do the research articles have adequate power. It should be noted, however, that large effect sizes are not generally encountered in behavioral research. What is more important, however, is that the power of the test in the studies reviewed was not determined by design but rather by default. Moreover, this picture is probably favorably biased due to the selectivity in accepting for publication articles that find significant results. Thus, in research in general, power is probably even lower than indicated in surveys such as this.

Summary and Recommendations

There are several approaches to increasing power:

(1) The most obvious immediate remedy to the problem of low power is to increase the number of subjects. Other factors held constant, this will result in increased power.

(2) Increasing the level of alpha will result in more power, however, "alpha should not be set thoughtlessly, but should reflect a balance in Type I-Type II error considerations" (Hopkins, 1973, p. 106).

(3) Research may be designed to increase the size of the effect under study rather than passively attempting to detect whatever effect is obtained, regardless of how small the effect is (Cohen, 1973).

(4) Other things being equal, a test of a directional hypothesis has more power than a nondirectional one. Directional tests, however, should be used judiciously (Cohen, 1969).

(5) Another important aspect is the reliability of the measures. As currently used, power analysis for the most part assumes high reliability. To the extent that the measures are unreliable, power will be less than that expected under optimal conditions (Cleary & Linn, 1969). Needless to say, high reliabilities are not the norm in behavioral research. This would lend further support to the assertion that in reality power is probably lower than what was found to be in these surveys.

In sum, the present survey indicated a serious shortcoming in the design of research in special education. In our continued efforts to upgrade such research it is important that considerations of statistical power analysis become an integral part of the training of researchers in our field, as well as one of the criteria for the evaluation of research reports submitted for publication.

References

Brewer, J. K. On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 1972, 9, 391-401.

Cleary, T. A., & Linn, R. L. Errors of measurement and the power of a statistical test. British Journal of Mathematical and Statistical Psychology, 1969, 22, 49-55.

Cohen, J. The statistical power of abnormal-social psychological research: a review. Journal of Abnormal and Social Psychology, 1962, 65, 145-153.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Cohen, J. Statistical power analysis and research results. American Educational Research Journal, 1973, 10, 225-230.

Hopkins, K. Preventing the number-one misinterpretations of behavioral research, or how to increase statistical power. Journal of Special Education, 1973, 7, 103-107.

Miller, J., & Knapp, T. The importance of statistical power in educational research. Bloomington, Ind.: Phi Delta Kappa, 1971.

Footnote

[1]While *r* is not a test of significance but a measure of association, it was decided to treat it in a separate category for the purpose of distinguishing studies reporting correlations from those focusing on mean differences, as well as for comparisons with other surveys available in the literature.

Table 1

Frequency Distributions of the Power of Statistical Tests in Exceptional Children, V39 & V40

	Effect Size								
	Small			Medium			Large		
Power[a]	t	F	r	t	F	r	t	F	r
91-100	1		1	1	2	1	2	9	5
81-90					1		1	1	2
71-80				2	1	4			
61-70					3		6		2
51-60					2		3	1	
41-50				1	1	1	3	1	
31-40		2		5		1	1	1	
21-30	1			5	1	2		2	
11-20	2	8	4	2	4				
0-10	12	5	4						
Number of Articles	16	15	9	16	15	9	16	15	9
Median	09	12	14	31	55	73	64	93	96

[a]Decimal points omitted.